DOCUMENT RESUME

ED 473 811 TM 034 817

AUTHOR	Arnemann, Kelly
TITLE	A Review of the Panoply of Effect Size Choices.
PUB DATE	2003-02-00
NOTE	22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, February 13-15, 2003).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Effect Size; *Meta Analysis; *Research Reports; Scholarly Journals

ABSTRACT

Some 23 journals in educational psychology and related fields, including two organizational "flagship" journals with circulations over 50,000, now "require" effect size reporting. This paper reviews some of the many effect size choices available to researchers. Effect size measures can be grouped into measures of association strength and measures of standardized mean difference. Another area of interest is that of effect sizes that address group overlap. Because there is no "one-size-fits-all," researchers must choose the best index for each study. The ability to perform meta-analysis and replication of research is determined by the inclusion of the effect size as a supplemental statistic. (Contains 1 table and 49 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 473 811

Running head: PANOPLY OF EFFECT SIZE CHOICES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
K. Arnemann
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

A Review of the Panoply of Effect Size Choices

Kelly Arnemann

Texas A&M University 77843-4225

TM034817

BEST COPY AVAILABLE

Paper presented at the annual meeting of the
Southwest Educational Research Association,
San Antonio, February 13-15, 2003.

Abstract

Some 23 journals, including two organizational "flagship" journals with circulations both greater than 50,000, now "require" effect size reporting. The present paper will review some of the numerous effect size choices available to researchers.

A Review of the Panoply of Effect Size Choices

The APA Task Force on Statistical Inference emphasized that effect sizes (e.g. Cohen's d, omega squared, eta squared) should "always" be reported with p values, and that "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (p. 599, emphasis added). And the new fifth edition of the APA (2001) Publication Manual emphasizes that:

> It is almost always necessary to include some index of effect size or strength of relationship...The general principle to be followed...is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect of relationship. (pp. 25-26, emphasis added)

Today, 23 journals require effect size reporting. Two of these journals have subscriptions **greater than 50,000**! For example, the guidelines for Exceptional Children now ask, "Have you addressed the practical significance of your findings using effect size indicators and/or narrative analyses?" (2000, 66(3), p. 416). And the Guidelines for Authors of the Journal of Counseling and Development now state, "Authors are expected to discuss the clinical significance of the results (one means to accomplish this is to report effect sizes" (2001, 79(2), p. 253). Two of these journals are organizational "flagship" journals (Council for Exceptional Children and American Counseling Association) of their respective associations.

Effect sizes are used as an alternative to or supplement for statistical significance tests, given the severe limits of statistical significance tests (cf. Cohen, 1994; Meehl, 1978; Schmidt, 1996; Thompson, 1996). There are various articles that explain different effect size choices (cf.

Cortina & Nouri, 2000; Kirk, 1996, in press; Olejnik & Algina, 2000; Rosenthal, 1994; Snyder & Lawson, 1993; Thompson 2002).

But there are 41 to 61 different effect size choices (Elmore & Rotou, 2001; Kirk, 1996)! And these do not even include the new group overlap I indices developed by Huberty and his colleagues (Hess, Olejnik & Huberty, 2001; Huberty & Holmes, 1983; Huberty & Lowman, 2000). Thus SERA members may appreciate an integrated review of some of the many available effect size choices, and especially the Huberty indices, now that more and more journals are requiring effect size reporting.

As of January 2003, the editorial policies of the following 23 journals require effect size reporting:

- Career Development Quarterly
- Contemporary Educational Psychology
- Early Childhood Research Quarterly
- Educational and Psychological Measurement
- Educational Technology Research & Development
- Exceptional Children
- Journal of Agricultural Education
- Journal of Applied Psychology
- Journal of Community Psychology
- Journal of Consulting & Clinical Psychology
- Journal of Counseling and Development
- Journal of Early Intervention
- Journal of Educational Psychology
- Journal of Educational and Psychological Consultation
- Journal of Experimental Education
- Journal of Experimental Psychology: Applied
- Journal of Learning Disabilities
- Journal of Personality Assessment
- Language Learning
- Measurement and Evaluation in Counseling and Development
- The Professional Educator
- Reading and Writing
- Research in the Schools

Definition of an Effect Size

An effect size is a name given to a family of indices that measure the magnitude of a treatment effect. It can be used to mean "the degree to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false" (Cohen, 1988). It also tells us to what degree the dependent variable can be controlled, predicted or explained by the independent variable(s) (Olejnik & Algina, 2000; Snyder & Lawson, 1993). Because there are many effect size choices and therefore no concept of "one-size fits all" (Thompson, 1999) the indices used for data analysis must be carefully chosen by the researcher so as to be deemed appropriate for that specific research project. This is not a new concept. Ronald Fisher (1925) proposed that researchers supplement the significance test in analysis of variance with the correlation ratio eta, which measures the strength of association between the independent and dependent variables. Kirk (1996) uses the term "effect magnitude" to refer to the supplemental measures that quantitative psychologists proffer. Effect size measures are also used in meta-analysis studies in order to summarize the findings from a specific area of research.

Two Families of Effect Sizes

Measures of Association Strength

Effect sizes can be measured by using a wide array of formulas (Kirk, 1996), however, Rosenthal (1994) classified effect sizes into two families: the *r* family and the *d* family. The *r* family includes the Pearson product moment correlation coefficient as well as the various squared indices of *r* and *r*-type quantities. The *d* family includes mean differences and standardized mean difference indices (Elmore & Rotou, 2001). In 1990, Maxwell and Delaney used the terms 'measures of association strength' to describe the *r* family indices and "measures

of effect size" for the *d* family indices. Several different choices of effect sizes are discussed and illustrated below (Snyder & Lawson, 1993).

The *r* family includes the Pearson product-moment correlation coefficient, which is utilized in studies using bivariate correlation. Measurements in the *r* family are classified as "uncorrected" effect size and "corrected" effect size. Two measurements that compute uncorrected effect size measurements for strength of association are R squared (R^2) and eta squared (η^2). Studies using multiple regression procedures use the coefficient of determination, which is the obtained, squared multiple correlation, R squared (R^2). This coefficient expresses the proportion of variance in the dependent variable accounted for by the linear combination of independent variables (Elmore & Rotou, 2001).

In the analysis of variance (ANOVA) and analysis of covariance (ANCOVA) the measures of an effect size are measures of the degree of association between an effect (e.g., a main effect, an interaction, a linear contrast) and the dependent variable. In (ANOVA/ANCOVA) the effect size eta squared (η^2) is used. Computationally, R^2 and η^2 are the same.

$$R^2 \text{ and } \eta^2 = SS_{explained} / SS_{total}$$

Note. SS = Sum of Squares

There are also two measurements that compute corrected effect size measurements. These are omega squared (ω^2) and epsilon squared (Rosnow & Rosenthal, 1996; Thompson, 1996). The formulas for each are as follows:

$$\omega^2 = SS_{explained} - [(v\text{-}1) * MS_{error}] / SS_{total} + MS_{error}$$

$$\text{Epsilon}^2 = SS_{explained} - [(v\text{-}1) * MS_{error}] / SS_{total}$$

Note. SS = Sum of Squares, v = number of levels in a factor, MS_{error} = mean square error

The corrected effect size estimate for the R^2 ($\check{R}^2$) overestimation, and the factors that affect the size of R^2: (e.g. the ratio of the number of independent variables or predictors to the size of the sample and the value of R^2) was discussed by Pedhazur (1997). The "shrinkage" (p. 208) can be estimated by applying the following formula:

$$\check{R}^2 = 1 - (1-R^2) * [(N-1)/(N-k-1)]$$

Note. N = Population, k = number of groups

In addition to the concept of shrinkage relative to the population squared multiple correlation coefficient, Pedhazur (1997) was also concerned with the replication of findings with the statement regarding the use of crossvalidation "to determine how well a regression equation obtained from one sample performs in another sample from the same population" (p. 209).

Sampling error is the difference between corrected and uncorrected effect size estimates. The uncorrected effect size estimates show whether or not the sample results can reproduce the unexplained variance from the sample data. A positive bias in the uncorrected effect size estimate occurs when the researcher cannot partition out the sampling error variance (Cromwell, 2001). Thompson (1996) explained that corrected effect size measurements may be used to estimate and adjust for the positive bias associated with three study features: smaller sample sizes, smaller population effects and the use of multiple variables. According to Thompson (1997), positively biased effect size overstates the effects that would be found in either the population or in future samples. All uncorrected effect sizes are positively biased effect estimates, but are less biased if (a) sample size is large, (b) population effects are large, and (c) few measured variables are used.

Standardized Mean Difference

In 1969, Cohen introduced the concept of d, which is the difference between the population mean divided by the average population standard deviation.

$$d = M_1 - M_2 / \sigma_{pooled}$$

Note. M_1 = Mean of population 1, M_2 = Mean of population 2 σ_{pooled} = Average population standard deviation.

Cohen's contribution to the field has also had lasting impact because he included guidelines for determining the magnitude of d. It was also the first effect size to be labeled as such. Cohen divided the range of magnitude into small, medium and large effects (Kirk, 1996). According to Cohen (1992) a medium effect of 0.5 was visible to the naked eye of the observer and several surveys have found that 0.5 approximates the average size of an observed effect in various fields (Cooper & Findley, 1982; Haase, Waechter & Soloman, 1982; Sedlmeier & Gigerenzer, 1989). A small effect of 0.2 is noticeably smaller than the medium effect but not small enough to usually be considered trivial. A large effect of 0.8 is the same distance from the medium effect as it is from the small effect (Kirk, 1996).

The second mean difference to be discussed is that of Glass' delta (Δ). Glass (1976) defined the effect size difference between the experimental group and the control group means divided by the standard deviation of the control group:

$$\Delta = M_e - M_c / S_c$$

Note. M_e = mean of experimental group, M_c = mean of the control group, S_c = standard deviation of the control group.

Glass replaced Cohen's d division of the difference between population means by the average population standard deviation with the sample standard deviation of the control group because "he reasoned that if there were several experimental groups, pairwise pooling of the standard deviations would result in a different standard deviation for each experimental-control contrast.

Hence, the same difference between experimental and control means would result in different effect size values when the standard deviation of the contrasts differed" (Kirk, 1996, pp.750-751).

A third measure of mean differences is Hedges g. Hedges (1981) pooled the standard deviations of the experimental group with the control group in order to have one standard deviation for all contrasts.

$$g = M_e - M_c / S_{pooled}$$

Note. M_e = mean of experimental group, M_c = mean of the control group, S_{pooled} = pooled standard deviation of the experimental group with the control group.

Cohen's d Glass' Δ and Hedges g are relevant when using a t-test. The main difference between the three formulas is found in the denominator.

The effect sizes in the mean differences and the Pearson r can also be transformed into each other's metrics (Thompson, 2000). Several examples will follow:

Cohen's d can be converted to an r using Cohen's (1988, p. 23) formula #2.2.6:

$$r = d / v[d^2 + 4].$$

Or r can be converted to d using Friedman's (1968, p. 246) formula #6:

$$d = [2(r)] / [v(1 - r^2)] \text{ (Thompson, 2000)}$$

or, d can also be computed from the value of the t-test of the difference between two groups

$$d = 2t / v(df) \qquad \text{or} \qquad d = t(n_1 + n_2) / [v(df) * v(n_1 * n_2)].$$

Note. df = degrees of freedom for t-test, n = number of cases for each group. The formula with the n's should be when the n's are *not* equal. The formula without the n's should be used when the n's are equal.

Cohen's d can also be computed from r the effect size correlation:

$$d = 2r / v(1-r^2).$$

Cohen's d can also be computed from Hedge's g:

$$d = g^{*}v(N / df).$$

Hedge's g can also be computed from the value of the t-test of the differences between groups:

$$g = 2t / vN \qquad \text{or} \qquad g = t^{*}v(n_1 + n_2) / v(n_1{}^{*}n_2).$$

Note. The formula with N is used when case numbers are equal. The formula with n should be used when case numbers are *not* equal.

Hedge's g can also be computed from r, the effect size correlation:

$$g = [r / v(1-r^2)] / v[df(n_1 + n_2) / (n_1{}^{*}n_2)].$$

The above formulas (Rosnow & Rosenthal, 1991) show the interplay of each of these effect size choices. It is up to the astute researcher to make the appropriate choices for the study being performed.

Group Overlap Indices

Another area of interest is that of effect sizes which address group overlap. Cohen has interpreted effect sizes in terms of the percentage of non-overlap between the treated group's scores and the untreated group. An effect size of 0.0 indicates that the distribution of scores for the treated group overlaps completely. The two groups are identical. An effect size of 0.8 indicates that a non-overlap of 47.4% (or an overlap of 52.6%). And an effect size of 1.7 indicates a non-overlap of 75.4% (or an overlap of 24.6%) in the two distributions. Please see Table 1 for this information (Cohen, 1988). The concept of group overlap will now be discussed.

The use of the overlap of two distributions of outcome scores as an effect size may make sense to some researchers (Huberty & Lowman, 2000). Tilton (1937) "suggested that the amount of group overlap be considered (in two-group univariate mean comparisons) in determining whether two means are significantly different" (Huberty, 2002, p. 232). Thirty years ago Alf and Abrahams (1968), Cohen (1969, p.10), Elster and Dunnette (1971), and Levy (1967) related

overlap to two-group mean difference testing. The specific instance of an *I*-like index was also suggested more than 30 years ago by Michael (1966). Group overlap was also revisited 15 years ago by Huberty and Holmes (1983), and Preece (1983) (Huberty, 2002). Of particular interest in the present paper is how Huberty and his colleagues perceive group overlap in the two-group and multiple outcome variable context. The improvement-over-chance classification (*I*) is what will now be discussed (Huberty, 2002). According to Huberty and Lowman (2000), the *I* index can be used for univariate, multivariate, homogeneous, heterogeneous, or any combination of the above research situations.

It must be noted that effect sizes used in standardized mean comparisons are restricted to the conditions of variance homogeneity. A good assessment approach is to use a univariate group membership prediction (or classification) rule (Huberty & Lowman, 2000).

A linear rule may be used if it is a univariate case and the variances can be assumed to be equal. Using this rule, the sample variances are pooled in order to compute the posterior probability estimates of group membership: $P(g / X_i)$. The estimates reflect the probability that the *i*th unit will belong to the g population given X_i as an observed score. To find the linear classification rule, the following formula can be used:

$$P(g / X_i) = q_g * \exp((-1/2)D^2_{ig}) / \sum_{g'=1}^{k} q_{g'} * \exp((-1/2)D^2_{ig'}).$$

Note. D^2_{ig} = Mahalanobis squared distance of unit i from the mean of group g (where X_g is the mean of group g); $s_{i'}$ = the pooled variance on the predictor variable; q_g = probability that any unit is a member of population g (Hess et al., 2001).

If population variances cann t be assumed to be equal, then the use of a quadratic classification rule would be required. The formula to obtain a quadratic rule is:

$$P(g / X_i) = q_g * s^{-1/2}_{g} * \exp((-1/2)D^2_{ig}) / \sum_{g'=1}^{k} q_{g'} * s^{-1/2}_{g'} * \exp((-1/2)D^2_{ig})$$

Note. The quadratic rule uses separate variance s_g (Hess et al., 2001).

Group overlap can then be assessed by using a prediction of group assignment by using predictive discriminant analysis (PDA) and logistic regression analysis (LRA) for the two-group comparison (Hess, Olejnik & Huberty, 2001). Three judgments must be made in creating classification rules: 1) determination of the normality of score distribution, 2) assessment of the equality of the two outcome-variable variances, and 3) the estimation of prior probabilities of group membership (as it relates to the sum to unity and relative sizes of the two populations). For a discussion of these three judgments review Huberty (1994, chap.4) (Huberty & Lowman, 2000).

Once the form of the rule is selected by taking the above three classification judgments into consideration, the method used to estimate group overlap must be selected. In order to determine group overlap using PDA, a group membership classification *error rate* must be calculated. The complement to the error rate, which is known as a *hit rate*, will be considered for the assessment of the group overlap (Huberty & Lowman, 2000). Huberty (1994) recommended using an external classification analysis in order to determine the hit rate. The classification rule for an external analysis is determined on one set of units, which is then used to classify the other sets of units in the analysis. (Hess et al., 2001). A hit rate estimate may be reached by using an external approach termed *leave-one-out* (L-O-O) by Huberty (1994, pp. 88-93) (Huberty & Lowman, 2000). The L-O-O method is also similar to the jackknife estimator (see Huberty, 1994). The L-O-O method will yield an acceptable point estimate of the hit rate because it will count the correctly classified units and serve as a good representation of group overlap (Hess et al., 2001).

An across-group hit rate estimate is a reasonable representation of group overlap. An interval estimate also needs to be established. To define an interval estimate, the meaning of "chance" for each particular study must be clarified. One interpretation of chance is based on the *proportional chance criterion*. With this interpretation, the chance frequency of hits for group g is: $e_g = q_g * n_g$.

Note. q_g = estimated prior probability for group g, n_g = number of analysis units in group g.

The across-group chance frequency of hits is: $e = \sum_{g=1}^{k} e_g$.

Note. k = the number of groups.

The across-group hit rate is $H_e = e / N$ and H_o is the notation for the across-group hit rate. Another interpretation of chance is the maximum chance criterion and it would be appropriate with a two-group situation with prior probabilities that are very different. This formula would be: $He = \max(q_1, q_2)$. Whether the proportional chance criterion or the maximum chance criterion is used is left to the judgment of the researcher (Huberty & Lowman, 2000).

Because a hit rate point estimate may not be an adequate effect size index, there is a need for an improvement-over-chance index. Huberty (1994, p.107) suggested the following index:

$$I = (1-H_e) - (1-H_o) / 1 - H_e$$

$$= H_o - H_e / 1 - H_e$$

The *I* index can be used to answer the question: "To what extent is the group distribution overlap more than what may be expected by chance (sampling variability)?"(Huberty & Lowman, 2000, p. 547). In order to use an *I* index as an effect size estimate there are two judgments that must be followed. The first judgment is with regard to the prior probabilities of group membership and that the proportions reflect the relative sizes of the populations of the groups involved in the

comparison, and the second judgment is that of using the appropriate interpretation of chance (Huberty & Lowman, 2000).

Some preliminary supporting evidence is provided by Huberty and Lowman (2000) for the use of *I* as a measure of effect size for univariate and multivariate group comparisons. The data set used in Huberty and Lowman (2000) was that of the BISBEY data set included in Huberty (1994) where N=153. Using this data, they compared the *I* index to F, η^2, and the point biserial correlation (p_{br}). In groups numbering more than 2, with homogeneity of variance conditions met, the relationship between F and *I* was .93, between η^2 and I was .97. In the two-group comparison situation, the relationship between p_{br} and *I* was .90. In non-homogeneous variance cases, Huberty and Lowman (2000) compared *I* to an adjusted *F* values (or *J* values) by utilizing the James second order test (Oshima & Algina, 1992). Getting the *I* values by use of the quadratic rule, the correlation between *J* and *I* values was found to be .89. The high correlations of these preliminary analyses are what led Huberty and Lowman (2000) to conclude that the *I* index could be used in univariate, multivariate, homogeneous, heterogeneous and any combination of contrasts deemed appropriate to the researcher (Hess et al., 2001).

Another method of two-group classification is the logistic regression analysis (LRA). This regression analysis models the dichotomous variable's nonlinear probabilistic function (Fan & Wang, 1999). In the two-group situation, given a dichotomous outcome variable Y and a single continuous predictor variable X, the posterior probability of membership in the target group (e.g. Group 1) is modeled by the logistic function: $Y = e^{\beta' x} / 1 + e^{\beta' x}$. Assuming there is only one predictor in the above equation, $\beta'X = \beta_o + \beta_1 * X_1$ and Y is the predicted posterior probability of belonging to the target group (Group 1). After the use of the logistic regression model is established, it can be used to obtain the hit rate. The process of

getting an observed hit rate from this point is simple: classify X_i into the other group if the predicted posterior probability of the observation for that group is small or into the target group (Group 1) if the predicted probability is large. The determination of cutoff points above which X_i is placed into the target group and below which point X_i is placed in the other group remains problematic. Each specific cutoff value is based on the size of the population being researched (Hess et al., 2001).

It has been shown that LRA can also assess group overlap between two population distributions based upon the computation of estimated hit rates and the subsequent *I* values that result. Future research can address under which conditions of variance heterogeneity would a researcher use quadratic PDA over LRA as the method for computing *I* (Hess et al., 2001).

Practical limitations that Hess et al. (2001) has stressed when using the *I* index are now discussed. With the particular use of PDA, depending on the distribution shape and ratio of the variances, theoretical values of *I* will be different. When using quadratic PDA as variance patterns become more extreme, *I* values are less differentiated in terms of small, medium and large. Because social science data collection typically occurs in less than ideal conditions, researchers cannot make attempts to make strict qualitative judgments based on sample estimates of the *I* regardless of which method is used to compute it. Hess et al. (2001) stressed that under any conditions that are considered less than ideal, the *I* index from any data set must be interpreted with caution, including those times when the suggested intervals from the Hess et al. (2001) study are being used. It is the conclusion of Hess et al. (2001) that LRA should be used in conjunction with the *I* index for the following reasons: 1) LRA does not require a test of variance equality, and 2) the logistic regression-based hit rates can be obtained from popular statistical software packages (e.g., SPSS and SAS). They also suggest that *I* will optimally perform with

improved precision and accuracy when used with large sample sizes (N=300) if the researcher can maintain an equal *n* ratio when the populations of the two groups are equal.

Conclusion

In summary, there are many choices for effect sizes. Cohen's d, Glass's Δ and Hedges g are popular choices, as are η^2 and ω^2. The fact that the formulas can be converted into each other's metrics only further supports the necessity of effect sizes as supplemental statistics. The introduction and review in this paper of the Huberty *I* index brings up the idea of generalizability across data analysis situations. "Conceptually, the *I* index is judged to be a reasonable index of group overlap and fairly straightforward in understanding" (Huberty & Lowman, 2000, p. 559). During further research on the *I* index, Hess et al. (2001) found that the use of LRA will benefit researchers the most if they are able to maintain the *n* ratio when the two populations are equal in size. Because there is no concept of "one-size fits all" (Thompson, 1999), it remains the choice of researchers to choose the best index for their particular work. The ability to perform meta-analyses and replication of the research is determined by the inclusion of the effect size as a supplemental statistic.

References

Alf, E. & Abrahams, N.M. (1968). Relationship between percent of overlap and measures of correlation. Educational and Psychological Measurement, 28, 779-792.

American Psychological Association. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC: Author

Cohen, J. (1994). The earth is round (p< .05). American Psychologist, 49, 997-1003.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cohen, J. (1988). Statistical power analysis of the behavioral sciences (2nd ed.). New York: Academic Press

Cohen, J. (1969). Statistical power analysis of the behavioral sciences. New York: Academic Press

Cooper, H. & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. Personality and Social Psychology Bulletin, 8, 168-173.

Cortina, J.M. & Nouri, H. (2000). Effect size for ANOVA designs. Thousand Oaks, CA: Sage

Cromwell, S. (2001, February). An introductory summary of various effect size choices. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans. (ERIC Documentation Reproduction Service No. TM 032 357)

Elmore, P. & Rotou, O. (2001, April). A primer on basic effect size concepts. Paper presented at the annual meeting of the American Educational Research Association, Seattle. (ERIC Documentation Reproduction Service No. ED 453 260)

Elster, R.S. & Dunnette, M.D. (1971). The robustness of Tilton's measure of overlap. Educational and Psychological Measurement, 31, 685-697.

Fan, X. & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. Journal of Experimental Education, 67, 265-286.

Fisher, R.A. (1925). Statistical methods for research workers. Edinburgh, UK: Oliver & Boyd.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. Psychological Bulletin, 70, 245-251.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Haase, R.F., Waechter, D.M. & Solomon, G.S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. Journal of Counseling Psychology, 29, 58-65.

Hedges, L.V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hess, B., Olejnik, S. & Huberty, C.J. (2001). The efficacy of two improvement-over-chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. Educational and Psychological Measurement, 61, 909-936.

Huberty, C.J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62, 227-240.

Huberty, C.J. (1994). Applied discriminant analysis. New York: John Wiley.

Huberty, C.J. & Holmes, S.E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15-26.

Huberty, C.J. & Lowman, L.L. (2000). Group overlap as a basis for effect size. Educational and Psychological Measurement, 60, 543-563.

Kirk, R. (1996). Experimental design: procedures for the behavioral sciences (3rd ed.). Pacific Grove, CA, Brooks/Cole.

Kirk, R. (1995). Practical significance: a concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Kirk, R.E. (in press). The importance of effect magnitude. In S.F. Davis (Ed.), Handbook of research methods in experimental psychology. Oxford, United Kingdom: Blackwell.

Levy, P. (1967). Substantive significance of significant differences between two groups. Psychological Bulletin, 67, 37-40.

Maxwell, S.E. & Delaney, H.D. (1990). Designing experiments and analyzing data: a model comparison perspective. Belmont, CA: Wadsworth.

Meehl, P.E. (1978). Theoretical risks and tabular astericks: Sir Karl, Sir Ronald, and slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Michael, W.B. (1966). An interpretation of the coefficients of predictive validity and of determination in terms of the proportions of correct inclusions or exclusions in cells of a four-fold table. Educational and Psychological Measurement, 26,419-425.

Olejnik, S. & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemporary Educational Psychology, 25, 241-286.

Oshima, T.C. & Algina, J. (1992). A SAS program for testing the hypothesis of the equal means under heteroscedasticity: James's second order test. Educational and Psychological Measurement, 52, 117-118.

Pedhazur, E.J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Preece, P.F.W. (1983). A measure of experimental effect size based on success rates. Educational and Psychological Measurement, 43, 763-766.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), The handbook of research synthesis, (pp. 231-244). New York: Russell Sage Foundation.

Rosnow, R.L. & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychological Methods, 1, 331-340.

Rosnow, R.L. & Rosenthal, R. (1991). Essentials of behavioral research: Methods and data analysis (2nd ed.). New York: McGraw-Hill.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-315.

Snyder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22(1), 2-5.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1997). Computing effect sizes. [on-line] Http://acs.tamu.edu/"bbt6147/effect.html

Thompson, B. & Snyder, P. (1998). Statistical significance and reliability analyses in recent JCD research articles. Journal of Counseling and Development, 76, 436-441.

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, Along with a primer on both effect size and the bootstrap. Invited address presented at the annual meeting of the American Educational Research Association, Montreal (ERIC Document Reproduction Service No. ED 429 110)

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70, 80-93.

Thompson, B. (2002). "Statistical," "practical," and "clinical": how many kinds of significance do counselors need to consider? Journal of Counseling and Development, 80, 64-71.

Thompson, B. & Kieffer, K.M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. Research in the Schools, 7(2), 3-10.

Tilton, J.W. (1937). The measurement of overlapping. Journal of Educational Psychology, 28, 656-662.

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: http:// www.apa.org/journals/amp/amp548594.html]

Table 1

Percentage of non-overlap (and overlap) according to Cohen's effect size standards.

Cohen's Standard	Effect Size	Percent of Non-overlap	Percent of Overlap
	2.0	81.1	18.9
	1.9	79.4	20.6
	1.8	77.4	22.6
	1.7	75.4	24.6
	1.6	73.1	26.9
	1.5	70.7	29.3
	1.4	68.1	31.9
	1.3	65.3	34.7
	1.2	62.2	37.8
	1.1	58.9	41.1
	1.0	55.4	44.6
	0.9	51.6	48.4
Large	0.8	47.4	52.6
	0.7	43.0	57.0
	0.6	38.2	61.8
Medium	0.5	33.0	67.0
	0.4	27.4	72.6
	0.3	21.3	78.7
Small	0.2	14.7	85.3
	0.1	7.7	92.3
	0.0	0.0	100.0

Note. Adapted from Cohen (1988).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

TM034817

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A REVIEW OF THE PANOPLY OF EFFECT SIZE CHOICES	
Author(s): KELLY ARNEMANN	
Corporate Source:	Publication Date: 2/14/03

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

XX

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: KELLY ARNEMANN	
Organization/Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone: 9798451335	FAX:
	E-Mail Address:	Date: 3/18/03

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)
PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.